Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 192 (2021) 5093–5103

Procedia Computer Science

www.elsevier.com/locate/procedia

25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Hybrid Deep Learning Vision-based Models for Human Object Interaction Detection by Knowledge Distillation

Oumaima Moutik[a], Smail Tigani*[a], Rachid Saadane[c], Abdellah Chehri[1]

[a]*Engineering Unit, Euro-Med Research Center, Euro-Med University. Fez, Morocco,*
[b]*Electrical Engineering Department, Hassania School of Public Labors, 20250 Casablanca, Morocco*
[c]*University of Quebec in Chicoutimi, Québec, Canada, G7H 2B1*

Abstract

People hope that computers can be in constant intelligence development. Just like humans, they can ”see” the world and ”recognize” a visual event. We propose an approach based on computer vision methods to recognize Human-Object interaction(HOI). The technique stands on aggregating significant contextual features Human-Object interactions and scene recognition. We design a branch architecture consisting of the main branch for HOI detection and a supplementary branch for scene recognition. We explore the deep learning models through the knowledge distillation method and the Cross Branch Integration mechanism for encoding models into graph neural network architecture. We construct a knowledge graph to merge between high-level context information. When trained collaboratively, those models allow computing efficiency, strong context knowledge.

Keywords: Computer Vision, Action Recognition, Human-Object Interactions (HOI), Scene Recognition, Deep Learning, Knowledge Distillation, Cross Branch Integration, Graph Neural Network, Knowledge Graph.

1. Introduction

Computer vision has been used to perform a variety of tasks. One of the hallmark tasks is Human object interaction (HOI) that requires a visual context analysis. Such clustering actions into one event may describe what is happening in a scene. However, one of the major obstacles faced is the lack of standardization of visual actions that share the same context from one human to another. To this end, it is still unable for machines to analyze and recognize human activities to a large extent [**?**]

Understanding dynamic scenes require large annotated datasets, which is a very time-consuming task. As a solution, researchers proposed methods to transfer knowledge of a network learned from the already annotated dataset to

* Corresponding author. Tel.: (+212661)97.29.00 ; Email : s.Tigani@outlook.com
E-mail address: s.Tigani@outlook.com

1877-0509
10.1016/j.procs.2021.09.287

a new network to accomplish a particular task. Intuitively, visual events is a highly semantic concept whereby various semantic cues [**?**] Human-object interaction detection is a relatively new task in the world of computer vision, and

Fig. 1. person read book

visual semantic information extraction [**?**]. When humans seek to interpret their environment, they observe other humans and how they interact with one another or objects. The first step in discovering a human-object interaction is to detect objects [**?**]. Object proposals recovered should contain at least one human for a human-object interaction to be present. Using these humans and object proposals, a model for solving this problem must then correctly identify a human-object interaction between the humans present and any objects in the image. For example, in figure 1, the activity is ”holding a book,” but due to the lack of standardization, the model may go wrong. For this reason, intelligent systems need to know the place or context, which helps understand what might have happened in the present. The scene has been addressed by ensemble techniques that combine different levels of semantics extracted from the images (e.g., recognized objects, global information, and context [**?**]).

Taking the figure which is located in ”the library”. Using deep learning and given that HOI detection is the main network and scene recognition is an auxiliary network, the human object interaction and scene recognition allow the tasks to work collaboratively via knowledge distillation [**?**]. By designing two-branch networks, a more principled approach learns rich context information for HOI detection without additional manual annotations. All of this information may lead us to determine the correct activity, which is ”person read book”.

Contribution To improve the HOI detection model, we propose a method that aggregates meaningful context features from another model with knowledge distillation. The model will learn HOI detection and scene recognition jointly. The contribution of this paper is summarised as follows:

- The method has two different branches the main branch for HOI detection. The second branch for scene recognition to help the primary model learn rich context information for HOI detection. The branches are trained collaboratively, allowing the model to recognize human activities efficiently explicitly.
- Knowledge graph for modeling the correlation between the two branches, the scene recognition is encoded into convolutional features by a Cross Branch Integration.

- Our approach outperformed the HOI detection model when it was by itself with a top-A accuracy by 9.7% on HOI detection method [**?**].

2. Related work

We review in this section the recent works on object detection, HOI detection, knowledge distillation, and graph neural network.

2.1. Object detection

As a longstanding, object detection represents a fundamental and challenging problem in the area of computer vision. The main goal of object detection is to determine whether there are any instances of objects from a particular category (person, cat, dog, etc.) in a given image or video [**?**].

Object detection forms the basis in computer vision for solving complex level vision analysis such as HOI detection. In recent years, deep learning techniques have achieved huge improvement for object detection. Many standard benchmark datasets are available like [**?**], [**?**] and various methods helps to detect objects such as YOLO [**?**] and SSD [**?**]. Those models predict boxes at three scales, extracting features from these scales using a similar concept to feature pyramid networks. Redmon in YOLO v3 [**?**] uses a hybrid approach to perform feature extraction, building on former YOLO v2 [**?**], Darknet-19 [**?**] and residual networks [**?**]. The new network, Darketnet-53, is significantly larger and has 53 convolutional layers.

In this work, we use the Feature Pyramid networks approach for object detection (FPN) [**?**] which is a method based on detecting objects at different scales. The procedure takes one single frame on input and outputs sized feature maps at multiple levels. We used Renset as backbone network [**?**] .

2.2. Human-Object Interaction Detection

Multiple elements can define a human activity [**?**]. Interpreting any human activity is still challenging to incorporate different external knowledge for recognition tasks accurately. An image may contain multiple humans performing the same interaction; for example, the same human simultaneously interacts with multiple objects (”sit on a couch and type on a laptop”). Several humans share the same interaction and object (”catch throw and ball”) or fine-grained interactions (”walk the horse,” or ”feed the horse”). For this reason, HOI detection must be done through different stages to analyze video content at a semantic level. In [**?**] authors tried to combine objects, scenes, and action recognition by using multiple instance learning models. In [**?**] authors explored the relation of the object and action by designing a discriminate classifier.

Reasoning over human interaction with objects (HOI) is essential for a complete understanding of the visual event. Human beings can use actions in various technological application domains such as intelligent surveillance systems, robotics, virtual reality, and soon on. However, understanding context knowledge is critical, and learning meaningful context knowledge is important to improve performance. Human-object interaction (HOI) detection strives to localize both the human and an object and the identification of complex interactions between them. The problem is challenging since it involves complex interactions that humans make with multiple objects, and things also interact with each other. Early work in HOI focus on Bayesian model [**?**], learned structured representations with spatial interaction [**?**] or based on handcrafted features e.g, SIFT [**?**] with object and human detectors. More recently, inspired by the notable success of deep learning and the availability of large-scale HOI datasets. Several deep learning-based HOI models were proposed. In [**?**], a Fast RCNN model for HOI recognition. In [**?**], the authors proposed a zero-shot learning model applied for addressing the long-tail problem in human-object recognition. A graph neural network like the graph parsing neural network (GPNN) [**?**], where the method infers a graph parsing that includes the HOI graph structure represented by adjacency matrices and node labels. The output explains a given scene with a graph structure, humans and objects are represented by nodes, and actions are defined as edges. for example, for ”person lick the knife,” the nodes are ”person” and ”knife,” and ”lick” is the edge.

2.3. Knowledge Distillation

In recent years, deep neural networks have been successful in both industry and academia, especially computer vision tasks [**?**]. The great success of deep learning is mainly due to its scalability to encode large-scale data and maneuver billions of model parameters. However, it is a challenge to deploy these cumbersome deep models on devices with limited resources, e.g., mobile phones and embedded devices, not only because of the high computational complexity but also the large storage requirements. To this end, a variety of model compression and acceleration techniques have been developed [**?**]. It aims at transferring knowledge acquired in one model (i.e., the teacher) to another model (i.e., the student) that is typically smaller.

A knowledge distillation system is composed of three key components: knowledge, distillation algorithm, and teacher-student architecture. The principal idea is that the student model mimics the teacher model to obtain a competitive or even superior performance. The method is similar to how human beings learn. This approach shows that softening the softmax predictions of a network by a high temperature conveys essential information, also called dark knowledge.

Knowledge distillation has been proposed for multi-modal action recognition [**?**].

2.4. Convolutional graph neural network

Convolution in GCNNs is the same operation in CNN [**?**]. It refers to multiplying the input neurons with a set of commonly known weights as filters or kernels. GCN performs similar operations where the model learns the features by inspecting neighboring nodes [**?**]. An image can be considered as a particular case of graphs where pixels are connected by adjacent pixels. Hence, in a 2D convolution (image), each pixel is taken as a node where the filter size determines neighbors. It takes the weighted average of pixel values of the node along with its neighbors. The forward propagation in the GCNN of a simple layer is:

$$H = \rho(\check{D}^{-1}\check{A}XW) \tag{1}$$

A is a matrix representing the edges or connection between the nodes in the forward propagation equation. The insertion of A in the forward pass equation enables the model to learn the feature representations based on nodes connectivity, $\check{A} = A + I$ is the adjacent matrix of the graph added self-loops, $\check{D}$ is its diagonal degree matrix with $\check{D}_i i = \sum j\check{A}_{ij}$, W is a matrix of trainable graph, X is the node information matrix $X \in R^{n*c}$, and ρ is a nonlinear activation function.

The graph convolution is separated into four steps, a linear feature transformation applied to the node information matrix by XW, mapping the c feature channels to c' channels to the next layer. The filter weights are shared among all vertices. The second step is $\check{A}XW$,hence, XW propagates node information to neighboring vertices as well as its self, where, $(\check{A}XW)_i = \sum \check{A}_{ij}(XW)_j = (XW)_i + \sum_{j\in T(i)}(XW)_j$, i.e, the $i^{(th)}$ row of the resulting matrix is the summation of $(XW)_i$ and $(XW)_j$ from $i's$ neighboring nodes. The third step tends to each row by multiplying $\check{D}_{ii}^{-1}$ to scale the features after graph convolution. The last step is to apply a point-wise nonlinear activation function ρ and output the graph convolution.

The graph convolution aggregates the node information in local neighborhoods to extract local substructure information. the results of each layer are transformed to the next layer in this form:

$$H^{t+1} = \rho(\check{D}^{-1}\check{A}H^t W^t) \tag{2}$$

Where $H^1 = X, H^t \in R^{n*c_t}$ is the output of the graph convolution layer. c_t is the number of output channels of layer, and $W^t \in R^{c_t * c_{t+1}}$ maps c_1 channels to c_{t+1} channels. After calculating the different graph convolution layers, another layer for concatenating the output $H^t, t = 1, ..., h$ horizontally to form a concatenated output,written as $H^{1:h} = [H^1, ..., H^h]$

, where h is the number of graph convolution layers and $H^1 : h \in R^{n*\sum_1^h c_t}$, In the concatenated output $H^{1:h}$, row represents the feature descriptor of a vertex,encoding its multi-scale local substructure information [**?**]. In our case, we apply one graph convolution layer, which has been demonstrated to be enough for modeling high-level context information.

3. Methodology

In this section, we describe the mechanism of the proposed method, which can learn the context knowledge of HOI detection and scene explicitly by training two different models jointly without extra manual annotations. We integrate a teacher network to distill the extra knowledge of scene information for additional supervision and a guide for HOI detection. We aggregate HOI and scene features by designing a knowledge Graph by CBI module.

3.1. The teacher network: Scene recognition

It's interesting to utilize pre-trained technologies separately while the annotations are highly expensive. We use a teacher network for scene recognition. Scene recognition has been addressed by ensemble techniques that combine different levels of semantics extracted from the images, for example, recognized objects, global information, and context at different scales [**?**]. We use Place365 [**?**] as a reference dataset, which is composed of 434 scenes which account for 98% of the type of scenes a person can encounter in the natural. The dataset contains 10 million images training set, validation set with 50 images per class, and test with 900 images per class.

Architectures trained on Place56	Validation set	Test Set
AlexNet	82.89%	82.75%
GoogleNet	83.88%	84.01%
VGG	84.91%	85.01%
ResNet	85.08%	85.07%

Authors in [**?**] propose different approaches to exploit the dataset at hand by training different CNN architectures like AlexNet. The performance of these architectures over the validation and test splits of the Places365 dataset is presented in Table 1. The authors experimented with the ResNet152 residual network architecture, fine-tuned over the Places365. This work achieved a top-5 accuracy of 85.08% and 85.07% on the validation and, respectively. The test set of the Places365 dataset, as shown in Table 2. For this reason, we use the Resnet architecture over the Place565 dataset in our model for scene recognition.

3.2. Main Network: Object-Human Interaction

HOI detection is an essential step towards detailed activity understanding. In this work, we followed the approach in [**?**], a detection method where interaction between human and objects are defined as a key point. The method directly detects interactions between human-objects pairs as a set of interaction points based on the human and object center points. The model learns to generate an interactive vector and then a grouping scheme that pairs the interaction with the correlate human and object bounding box predictions. The objective of this approach is to localize the agent (human) and object along with detecting. It comprises four steps: object detection, feature extraction, interaction generation, and interaction grouping. Consequently, the interaction point and vector and the detected human and object bounding boxes are input to the interaction grouping step for the final HOI triplet human, action, object prediction. In the following figure, we describe the overall HOI approach. First, For object detection, we employ the FPN object detector method to generate all the possible bounding box human and objects instances in a frame. Then, for feature extraction, we follow the same methodology of [**?**] by employing the Hourglass [**?**] network as the backbone network. The output of the Hourglass network is a feature map with size $H/S * W/S * D$, where H is the height and W is the width of the input frame, and D, S is the output channels and stride.

The output of the backbone network will be the input to the interaction generation module to produce the interaction point and interaction vector. Interaction point is defined as the center point of the action between a human-object pair

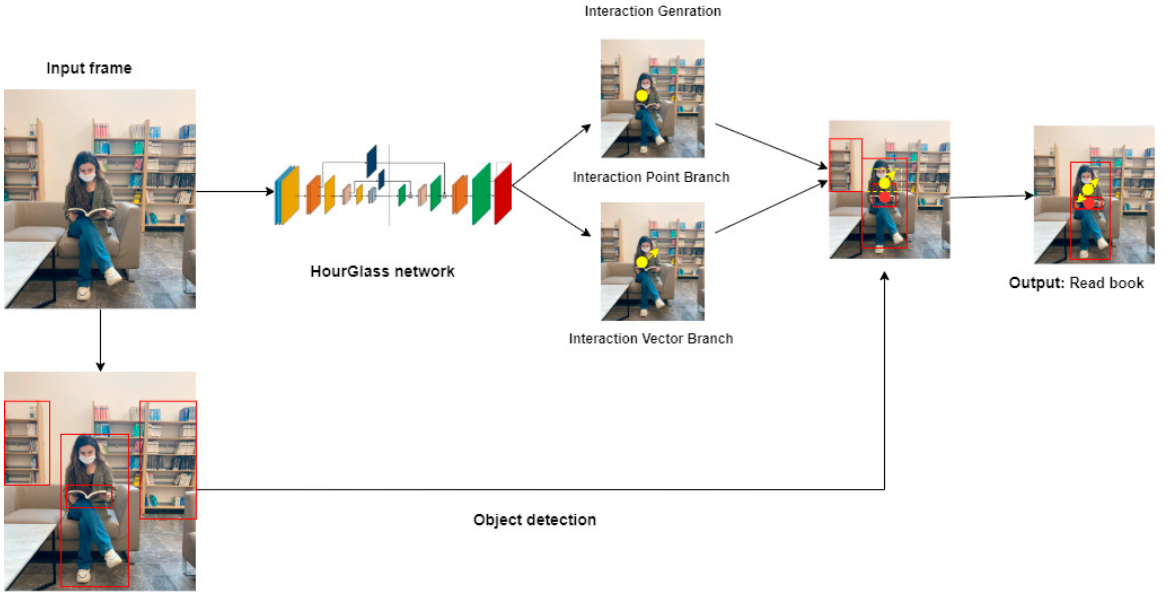
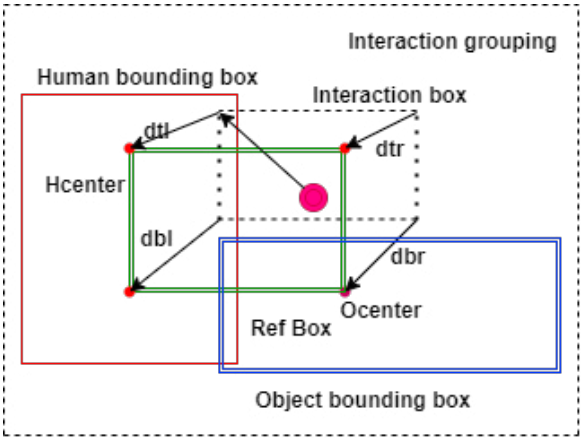


Fig. 2. the architecture of the HOI detection method. Workflow of the proposed HOI detection approach having a localization and an interaction prediction stage, we adopt a standard object detector (FPN [**?**]) to obtain human and object bounding-box predictions. Three steps of interaction prediction, (1) feature extraction, (2) interaction generation and (3) interaction grouping. The interaction generation contains two independent branches to produce interaction point and interaction vector, respectively. Interaction point and vector together with the recognized human and object bounding-box predictions are then inputted to the interaction grouping for final HOI predictions: human, action, object

.

and is the starting point of the interaction vector. Consequently, the interaction point and vector and the detected human and object bounding boxes are input to the interaction grouping step for the final HOI triplet human, action, object prediction as shown in the following.



hoi.png

Fig. 3. The procedure of interaction grouping scheme. It has three inputs: the human/object bounding-boxes from object detection branch, the interaction points from the interaction point branch and the interaction vector predicted by the interaction vector branch.

3.3. Knowledge Graph Mechanism

Our goal is to develop a feature integration approachable to aggregate context knowledge of different levels from the scene recognition and the HOi detection models. To do so, we followed the method in [**?**] based on two groups:

Knowledge distillation and Convolutional neural network. The process encodes the knowledge of scene recognition into HOI detection for more accuracy following the student-teacher architecture.

3.3.1. Knowledge distillation

This method aims to transfer knowledge between modalities, i.e., scene recognition and HOI detection. After training the scene recognition network separately, we transfer the knowledge from this model (the teacher network) to the HOI detection model (the student model). We use the features obtained by the student model for the HOI detection as gated modulation of the main HOI features via implementing element-wise multiplication on them like in the following equation (Equation 3).

$$f(x_a, x_b) = ReLU(W[\theta(x_a), \phi(x_b)]) \tag{3}$$

3.3.2. Knolwegde graph neural network

We obtain four groups of representations vectors of the same size; each ensemble contains N feature vectors corresponding to the number of frames. Following the method in [J'], we construct a knowledge graph to model the pair-wise correlation among the representations explicitly. the method is based on Graph convolutional network (GCN) to represent HOI detection and scene relationships. we construct the graph G graph V ,denoted the nodes $X = \left\{x_i^{human}, x_i^{action}, x_i^{object}, x_i^{scene}\right\}$,where $i \in N$ and $X \in R^d$, with d indicating the channel dimension of the last convolutional layer in the backbone. The graph G represents the pair-wise relationship among the nodes, The structure of our graph is fuly connected knowledge graph as illustrated in the figure 4.

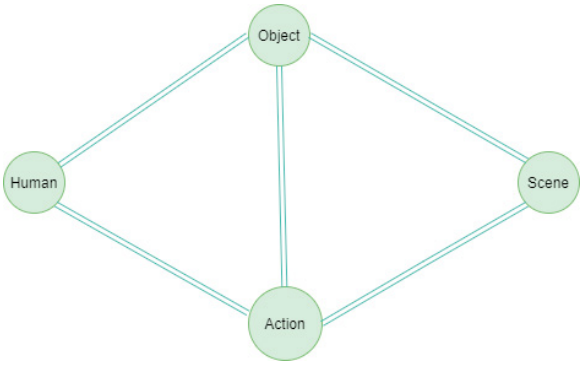


Fig. 4. The structure of the graph

.

The graph $G \in R^{N*N}$ represents the pair-wise relationship among the nodes, with edge G_{ab} indicates the relationship between node x_a and x_b, to compute the relations between nodes and to build the correlation between HOI detection and scene recognition, we adopt the relation module by concatenation [K'].

$$G_{ab} = \frac{e^{f(x_a, x_b)}}{\sum_{b=1}^{N} e^{f(x_a, x_b)}} \tag{4}$$

here $[\cdot, \cdot]$ denotes the concatenation operation, and W represents the learnable weight matrix that projects the concatenated vector into a scalar. To normalize the knowledge graph, we apply the softmax function: We applied the softmax function for implementing the normalization. Hence, the sum of all edges pointing to the same node must be normalized to 1 to cast the dot product into the Gaussian function for directly learning the relations. We apply GCN

[L'] on the constructed graph (figure) to aggregate g high-level semantic knowledge of cene into the HOI branch. We used graph convolution, which is enough for integrating rich high-level context information. The output of the GCN has the same size as the input X. We apply joint learning in order to train the tasks both collaboratively. While the teacher network provides the ground truth of scene recognition as pseudo labels for knowledge distillation.The multi task loss function is applied as:

$$L = \alpha_{HOI}[L_{human} + L_{action} + L_{object}] + \alpha_S L_{scene} \tag{5}$$

We set $\alpha_{HOI} = 1$ for HOI task, $\alpha_S = 0.01$ for the student network (scene).

4. Experiments

4.1. Training the Main Network

To main our study to an effective end, we conduct meticulous experiments on two large HOI datasets: V-COCO [**?**] and HICO-DET [**?**]. The V-COCO is a rich, varied dataset; It contains 2533 images for training, 2867 images of validation, and 4946 images for testing. To train the model, 5400 images representing the training and validation sets are required. Three action categories (cut, hit, eat) annotated with two types of targets (instrument and direct object) are considered. Additionally to 26 binary action labels and three classes (run, stand, walk) annotated with no interaction object are part of the Human instances in the V-COCO dataset. On the other hand, the HICO-DET dataset accommodates 38118 images for training and 9658 images for testing. Six hundred classes of different interactions annotating each human instance correspond to 80 object categories and 117 action verbs in this dataset. In our training phase, we followed the strategy in [**?**] for HOI detection by first using an FPN pre-trained object detector to initialize the framework to obtain the object's bounding box. Then we used the Hourglass network as a feature extractor for interaction prediction. The head network for the interaction point and interaction vector generation is randomly initialized. Yet, to achieve the optimization of the loss function during training, we engaged standard data augmentation techniques (random flip, random scaling) and Adam Optimizer [**?**]. At long last, to obtain final predictions during the test, we used flip augmentation.

4.2. Ablation study

In this part of the study, we will analyze each method and gather them little by little to determine the most efficient processes.

Method	Settings	top-1	gain
Baseline	HOI model	52.3	-
Knowedle distillation	Baseline+Scene recognition	57.3	+5
CBI	CBI+Baseline	59	1.7
KGCN	Knowledge graph embedding	60.1	+1.1
Our approach	KD+CBI+CBI	62	+1.9

4.2.1. Encoding Scene Recognition into HOI Detection by Knowledge Distillation

We use HOI detection based on points interaction [**?**] as a baseline, then we firstly include scene recognition into the HOI network by jointly learning the two tasks via knowledge distillation. The multitask learning with knowledge distillation outperforms the baseline (Table 2). When HOI detection is jointly learning scene recognition are jointly trained, the top-1 accuracy increases 5%.

4.2.2. CBI module

After encoding scene recognition into HOI detection by knowledge distillation, we applied the CBI module that enables the intermediate features exchange, unlike simple knowledge distillation, which allows a simple multitasking

learning. That is the main reason that justifies the enforcement of the learning ability of HOI by aggregating scene recognition into HOI detection. According to the experiment results (Table 2), when we applied CBI, we obtain higher accuracy with 1.7%.

4.2.3. Knowledge Graph Neural Network

The GCN method outperforms the two previous methods by boosting its performance using the aggregation of multiple branches and models the relation among human, action, objects, and scene representation. By applying dot embedded, the top-1 accuracy increases 1%.

4.2.4. Entire Framework

Finally, we combine all the previous components into the baseline (i.e., HOI detection), as shown in Table 1. We find that the top-1 accuracy has been boosted to 62%, while the baseline is 52%. This significant improvement of 9.7% on the HOI detection benchmark proves the effectiveness of our proposed framework.

5. Conclusion

Since HOI detection is the main determinator of human activity in video. We proposed in this work the improvement of this model by integrating scene recognition. We started with the knowledge distillation to merge the knowledge of scene recognition with the knowledge of the HOI model. To do that, we integrated the GCN network, which helped us learn relevant high-level semantic factors information.

In the future, we will improve our approach significantly by integrating further models into HOI detection, such as human posture and human skeleton.

acknowledgements

We would like to thank the anonymous referees for their valuable comments and helpful suggestions. Special thanks go to any one that improved the quality of the language and made this paper more readable.

Conflict of interest

The authors declare that they have no conflict of interest.

Appendices

Deep Learning Optimizers Overview

References

References

[1] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6, Antalya, August 2017. IEEE.

[2] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to Detect Human-Object Interactions. *arXiv:1702.05448 [cs]*, February 2018. arXiv: 1702.05448.

[3] Xiaojun Chen, Shengbin Jia, Ling Ding, Hong Shen, and Yang Xiang. SDT: An integrated model for open-world knowledge graph reasoning. *Expert Systems with Applications*, 162:113889, December 2020.

[4] Rishabh Dabral, Srijon Sarkar, Sai Praneeth Reddy, and Ganesh Ramakrishnan. Exploration of Spatial and Temporal Modeling Alternatives for HOI. page 10.

[5] V Delaitre, J Sivic, and I Laptev. Learning person-object interactions for action recognition in still images. page 10.

[6] Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. Knowledge Distillation: A Survey. *Int J Comput Vis*, March 2021. arXiv: 2006.05525.

[7] Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. Knowledge Distillation: A Survey. *Int J Comput Vis*, March 2021. arXiv: 2006.05525.

[8] Abhinav Gupta and Larry S. Davis. Objects in Action: An Approach for Combining Action Understanding and Object Perception. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, MN, USA, June 2007. IEEE.

[9] Shonosuke Harada, Hirotaka Akita, Masashi Tsubaki, Yukino Baba, Ichigaku Takigawa, Yoshihiro Yamanishi, and Hisashi Kashima. Dual graph convolutional neural network for predicting chemical networks. *BMC Bioinformatics*, 21(S3):94, April 2020.

[10] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *arXiv:1602.07360 [cs]*, November 2016. arXiv: 1602.07360.

[11] Nazli Ikizler-Cinbis and Stan Sclaroff. Object, Scene and Actions: Combining Multiple Features for Human Action Recognition. page 14.

[12] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale Residual Network for Image Super-Resolution. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, volume 11212, pages 527–542. Springer International Publishing, Cham, 2018. Series Title: Lecture Notes in Computer Science.

[13] Ruoyu Li, Sheng Wang, Feiyun Zhu, and Junzhou Huang. Adaptive Graph Convolutional Neural Networks. page 8.

[14] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2D-3D Joint Representation for Human-Object Interaction. page 10.

[15] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. *arXiv:1612.03144 [cs]*, April 2017. arXiv: 1612.03144.

[16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector. *arXiv:1512.02325 [cs]*, 9905:21–37, 2016. arXiv: 1512.02325.

[17] Arun Mallya and Svetlana Lazebnik. Learning Models for Actions and Person-Object Interactions with Transfer to Question Answering. *arXiv:1604.04808 [cs]*, July 2016. arXiv: 1604.04808.

[18] Alina Matei, Andreea Glavan, and Estefania Talavera. Deep learning for scene recognition from visual data: a survey. *arXiv:2007.01806 [cs]*, July 2020. arXiv: 2007.01806.

[19] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked Hourglass Networks for Human Pose Estimation. *arXiv:1603.06937 [cs]*, July 2016. arXiv: 1603.06937.

[20] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning Human-Object Interactions by Graph Parsing Neural Networks. *arXiv:1808.07962 [cs]*, August 2018. arXiv: 1808.07962.

[21] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning Human-Object Interactions by Graph Parsing Neural Networks. *arXiv:1808.07962 [cs]*, August 2018. arXiv: 1808.07962.

[22] Yijun Qian, Lijun Yu, Wenhe Liu, Guoliang Kang, and Alexander G. Hauptmann. Adaptive Feature Aggregation for Video Object Detection. In *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 143–147, Snowmass Village, CO, USA, March 2020. IEEE.

[23] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. *arXiv:1804.02767 [cs]*, April 2018. arXiv: 1804.02767.

[24] Shet Reshma Prakash and Paras Nath Singh. Object detection through region proposal based techniques. *Materials Today: Proceedings*, page S2214785321016746, March 2021.

[25] Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *arXiv:1706.01427 [cs]*, June 2017. arXiv: 1706.01427.

[26] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th international conference on Multimedia - MULTIMEDIA '07*, page 357, Augsburg, Germany, 2007. ACM Press.

[27] Mohammad Javad Shafiee, Brendan Chywl, Francis Li, and Alexander Wong. Fast YOLO: A Fast You Only Look Once System for Real-time Embedded Object Detection in Video. *arXiv:1709.05943 [cs]*, September 2017. arXiv: 1709.05943.

[28] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A Large-Scale, High-Quality Dataset for Object Detection. page 10.

[29] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling Human-Object Interaction Recognition Through Zero-Shot Learning. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1568–1576, Lake Tahoe, NV, March 2018. IEEE.

[30] Sasha Targ, Diogo Almeida, and Kevin Lyman. Resnet in Resnet: Generalizing Residual Architectures. *arXiv:1603.08029 [cs, stat]*, March 2016. arXiv: 1603.08029.

[31] Fida Mohammad Thoker and Juergen Gall. Cross-modal knowledge distillation for action recognition. *arXiv:1910.04641 [cs]*, October 2019. arXiv: 1910.04641.

[32] Yunong Tian, Guodong Yang, Zhe Wang, Hao Wang, En Li, and Zize Liang. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. *Computers and Electronics in Agriculture*, 157:417–426, February 2019.

[33] Torralba, Murphy, Freeman, and Rubin. Context-based vision system for place and object recognition. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 273–280 vol.1, Nice, France, 2003. IEEE.

[34] Shengquan Wang, Ang Li, Jiying Chen, Baoyu Zheng, Jiaxin Ji, and Li Xianglong. RSnet:An improvement for Darknet. page 8.

[35] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning Human-Object Interaction Detection Using Interaction Points. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4115–4124, Seattle, WA, USA, June 2020. IEEE.

[36] Zuxuan Wu, Yanwei Fu, Yu-Gang Jiang, and Leonid Sigal. Harnessing Object and Scene Semantics for Large-Scale Video Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3112–3121, Las Vegas, NV, USA, June 2016. IEEE.

[37] Shiwen Zhang, Sheng Guo, Limin Wang, Weilin Huang, and Matthew Scott. Knowledge Integration Networks for Action Recognition. *AAAI*, 34(07):12862–12869, April 2020.

[38] Zijun Zhang. Improved Adam Optimizer for Deep Neural Networks. In *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, pages 1–2, Banff, AB, Canada, June 2018. IEEE.

[39] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 14.

[40] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, June 2018.